Statistical Methods for Experimental Particle Physics

Theory and Lots of Examples

Thomas R. Junk
Fermilab

TRIUMF Summer Institute
July 20 - 31, 2009

Day 2: Hypothesis Testing
Confidence Intervals

Hypothesis Testing

- Simplest case: Deciding between two hypotheses. Typically called the *null* hypothesis H_0 and the *test* hypothesis H_1

- Can't we be even simpler and just test one hypothesis H_0?
  - Data are random -- if we don't have another explanation of the data, we'd be forced to call it a random fluctuation. Is this enough?
  - All models are wrong, but some are useful.

  H_0 may be broadly right but the predictions slightly flawed
  - Look at enough distributions and for sure you'll spot one that's mismodeled. A second hypothesis provides guidance of where to look.

- Popper: You can only prove models wrong, never prove one right.

- Proving one hypothesis wrong doesn't mean the proposed alternative must be right.

Frequentist Hypothesis Testing: Test Statistics and p-values

Step 1: Devise a quantity that depends on the observed data that ranks outcomes as being more signal-like or more background-like.

Called a test statistic. Simplest case: Searching for a new particle by counting events passing a selection requirement.

Expect *b* events in H_0, $s+b$ in H_1.

The event count n_{obs} is a good test statistic.

Step 2: Predict the distributions of the test statistic separately assuming:

H_0 is true

H_1 is true

(Two distributions. More on this later)

Frequentist Hypothesis Testing: Test Statistics and p-values

Step 3: Run the experiment, get observed value of test statistic.

Step 4: Compute p-value

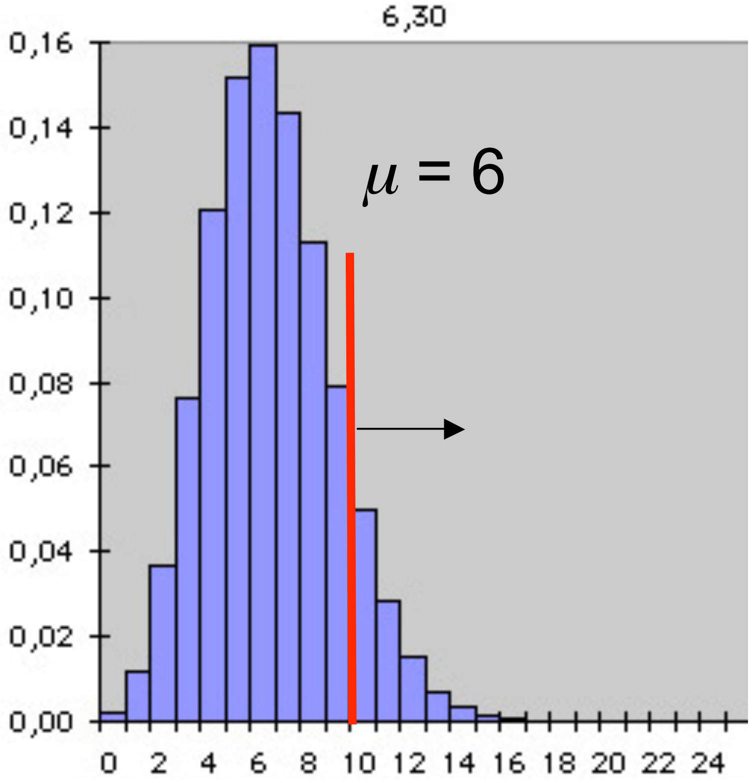
$p(n \geq n_{obs} | H_0)$

Example:

H_0: $b = \mu = 6$

$n_{obs} = 10$

p-value = 0.0839



A p-value is **not** the "probability H_0 is true"

But many often say that.

Common Standards of Evidence

Physicists like to talk about how many “sigma” a result corresponds to and generally have less feel for p-values.

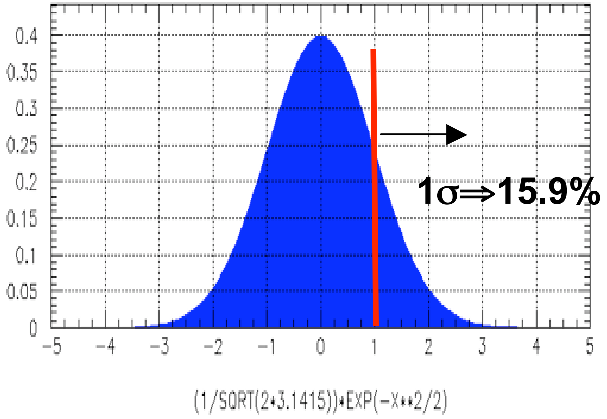
The number of “sigma” is called a “z-value” and is just a translation of a p-value using the integral of one tail of a Gaussian

```
Double_t zvalue = - TMath::NormQuantile(Double_t pvalue)
```

Folklore:
95% CL -- good for exclusion
3σ: “evidence”
5σ: “observation”
Some argue for a more subjective scale.

z-value (σ)	p-value
1.0	0.159
2.0	0.0228
3.0	0.00135
4.0	3.17E-5
5.0	2.87E-7

$$pvalue = \frac{\left(1 - erf\left(zvalue / \sqrt{2}\right)\right)}{2}$$



Tip: most physicists talk about p-values now but hardly use the term z-value

Sociological Issues

- Discovery is conventionally 5σ. In a Gaussian asymptotic case, that would correspond to a ±20% measurement.

- Less precise measurements are called “measurements” all the time

- We are used to measuring undiscovered particles and processes. In the case of a background-dominated search, it can take years to climb up the sensitivity curve and get an observation, while evidence, measurements, etc. proceed.

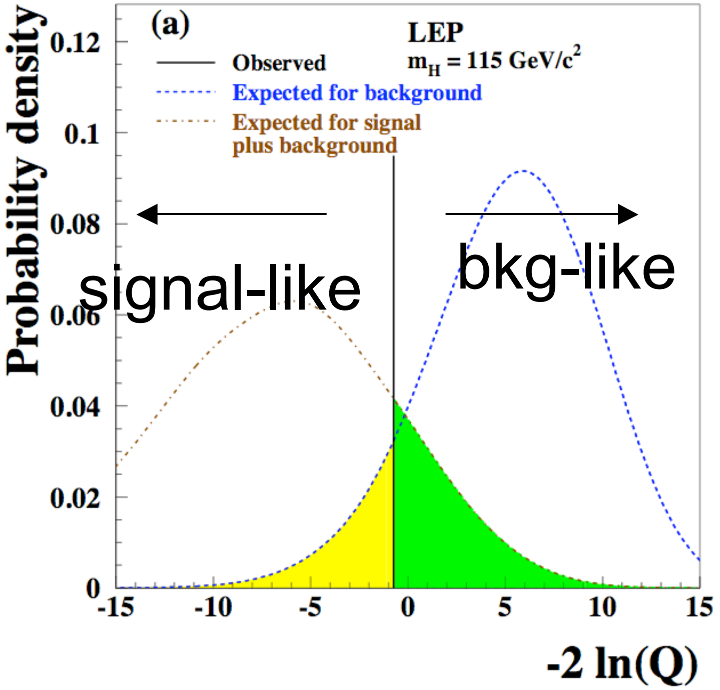
- Referees can be confused.

A More Sophisticated Test Statistic

What if you have two or more bins in your histogram? Not just a single counting experiment any more.

Still want to rank outcomes as more signal-like or less signal-like

Neyman-Pearson Lemma: The likelihood ratio is the “uniformly most powerful” test statistic



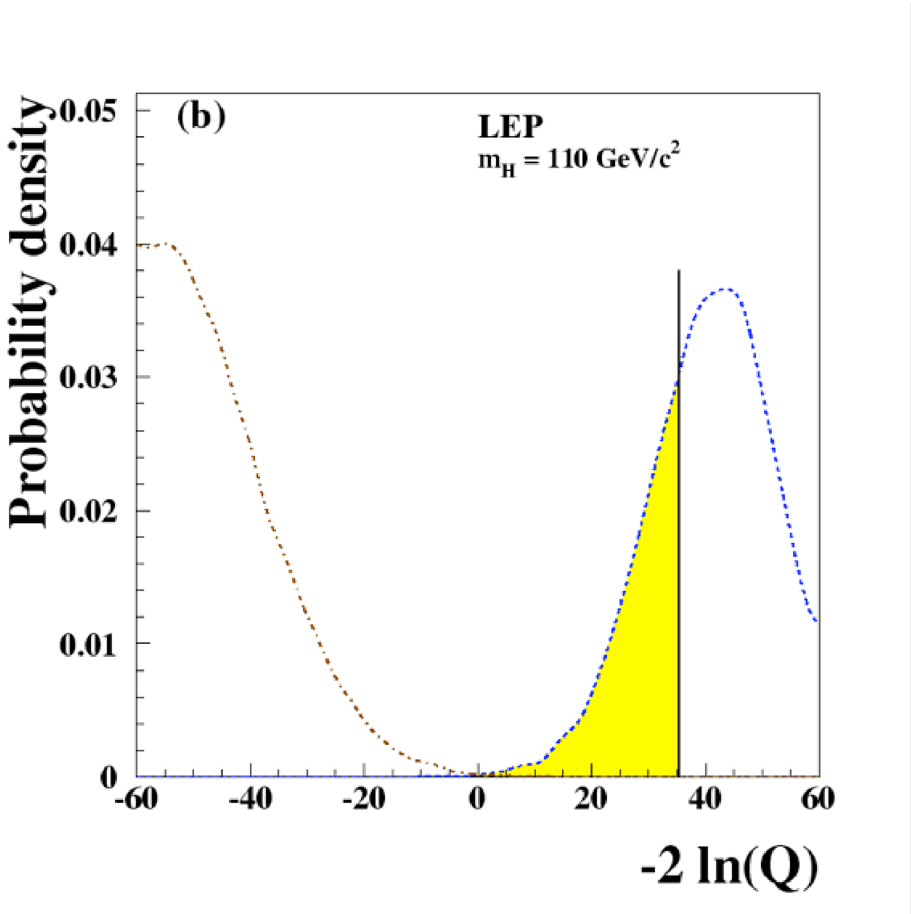
$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,H_1,\hat{\theta})}{L(\text{data}\,|\,H_0,\hat{\hat{\theta}})}\right)$$

yellow=p-value for ruling out *H_0.* Green= p-value for ruling out *H_1*

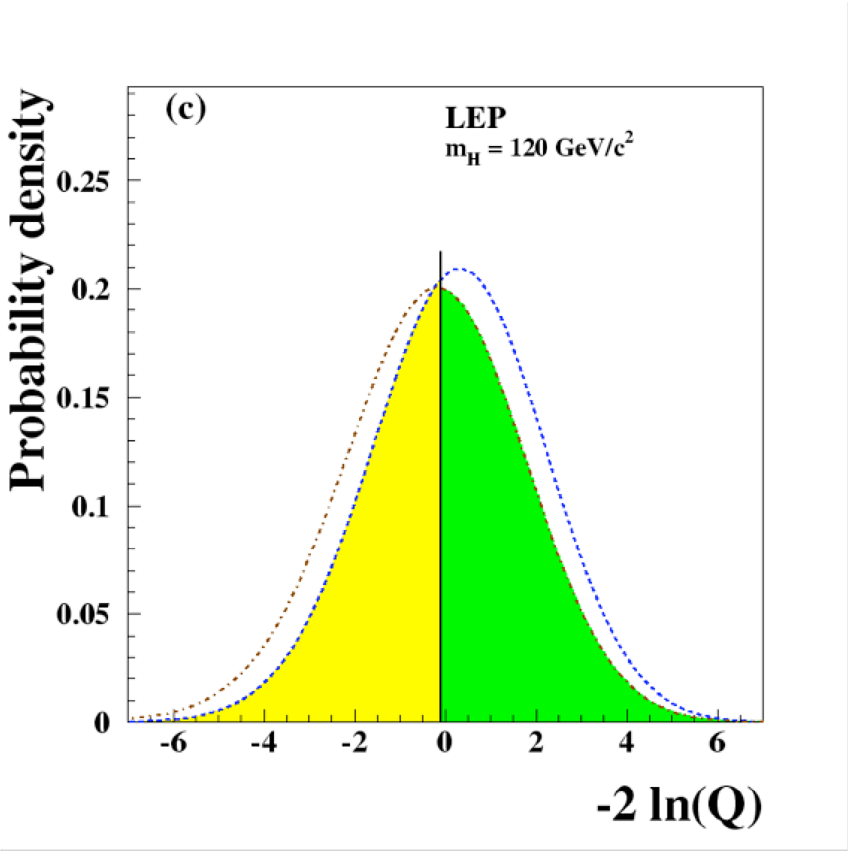
Acts like a difference of Chisquareds in the Gaussian limit

$$-2\ln Q \rightarrow \Delta\chi^2 = \chi^2(data\,|\,H_1) - \chi^2(data\,|\,H_0)$$

More Sensitivity or Less Sensitivity



signal p-value very small.
Signal ruled out.



Can make no statement regardless of experimental outcome.

What's with $\hat{\theta}$ and $\hat{\hat{\theta}}$?

A *simple hypothesis* is one for which the only free parameters are parameters of interest.

A *compound hypothesis* is less specific. It may have parameters whose values we are not particularly concerned about but which affects its predictions. These are called *nuisance parameters*, labeled θ.

Example: H_0=SM. H_1=MSSM. Both make predictions about what may be seen in an experiment. A nuisance parameter would be, for example, the b-tagging efficiency. It affects the predictions but in the end of the day we are really concerned about H_0 and H_1.

What's with $\hat{\theta}$ and $\hat{\hat{\theta}}$?

We parameterize our ignorance of the model predictions with nuisance parameters.

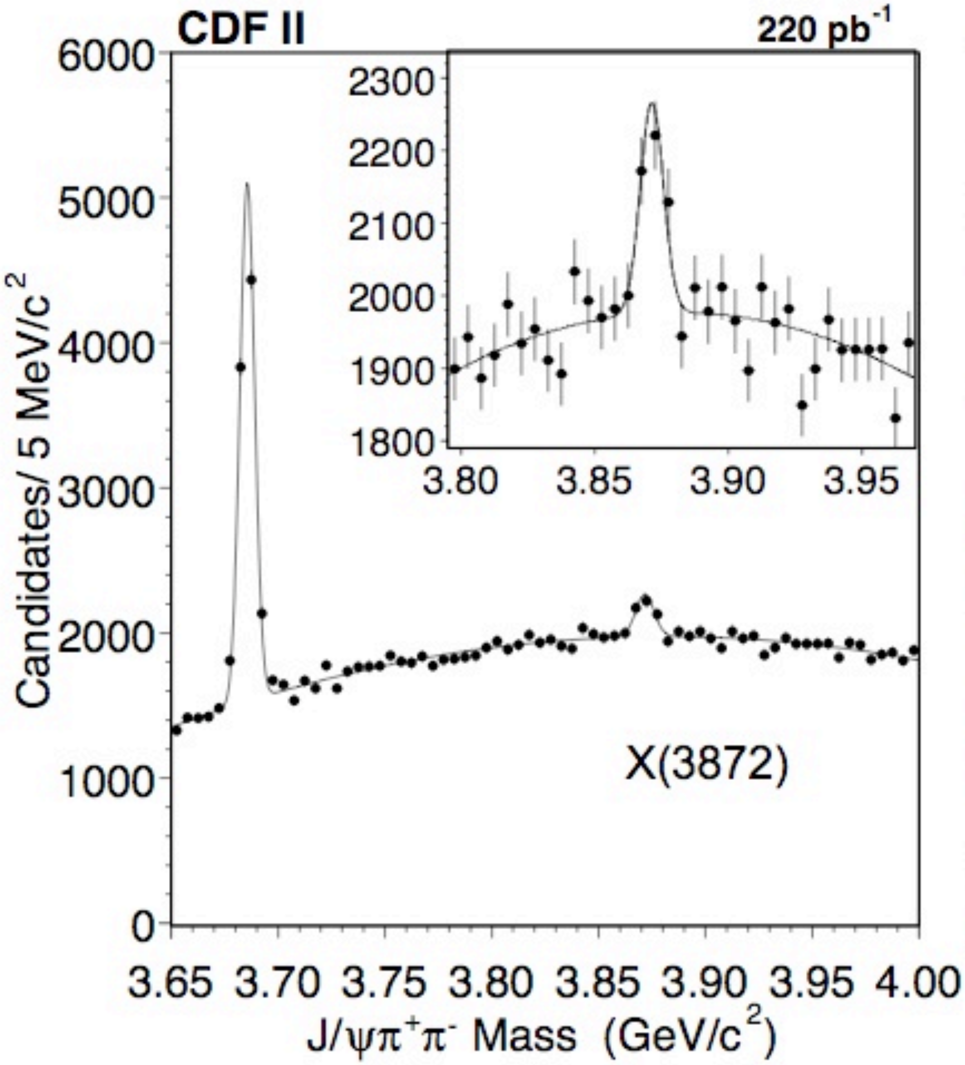
A model with a lot of uncertainty is hard to rule out!

-- either many nuisance parameters, or one parameter that has a big effect on its predictions and whose value cannot be determined in other ways

$\hat{\theta}$ maximizes L under H_1

$\hat{\hat{\theta}}$ maximizes L under H_0

The Traditional Solution to Large, Uncertain Backgrounds: Sideband Fits

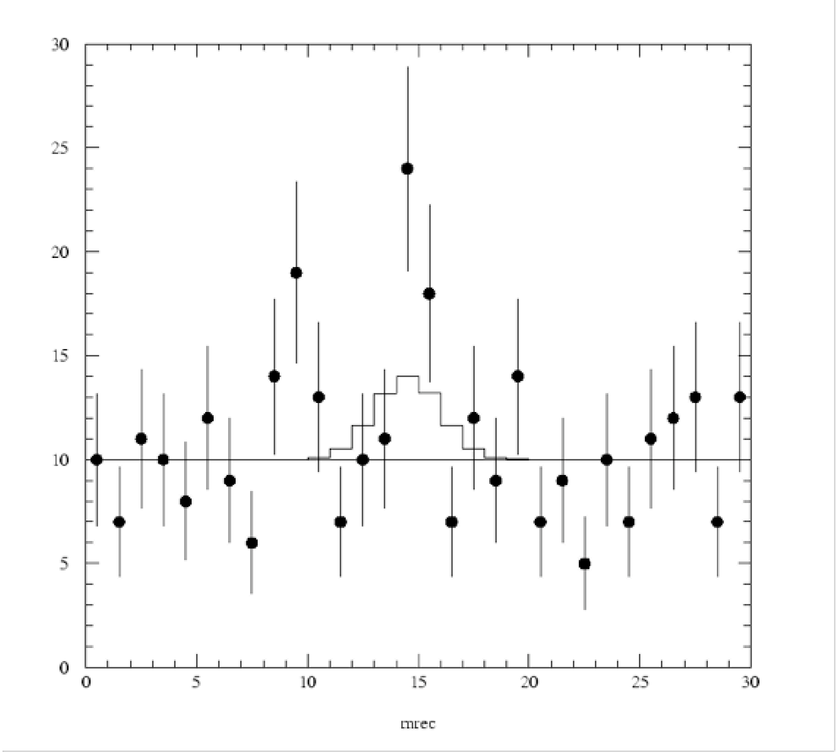


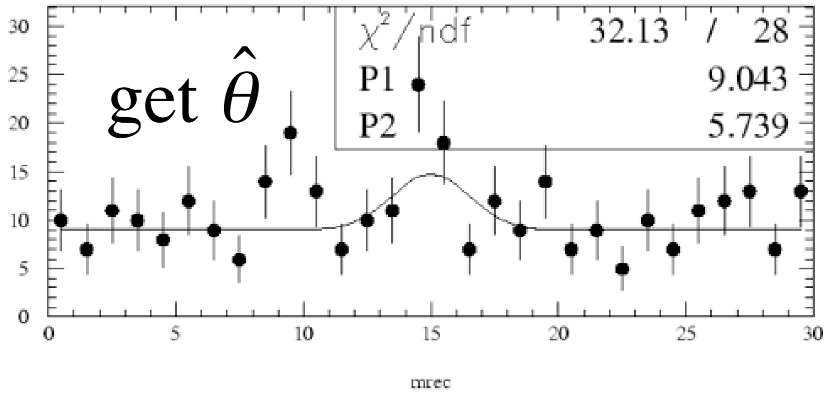
Guess a shape that fits the backgrounds, and fit it with a signal.

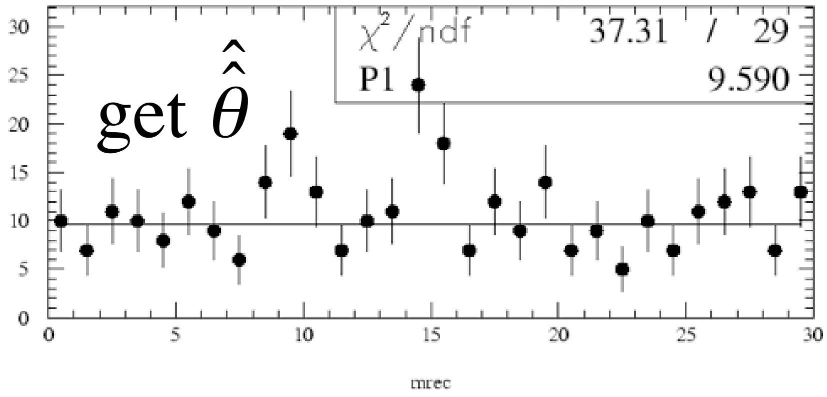
Fit twice! Once assuming H_0, once assuming H_1

Example: flat background, 30 bins, 10 bg/bin, Gaussian signal.
Run a pseudoexperiment (assuming s+b).

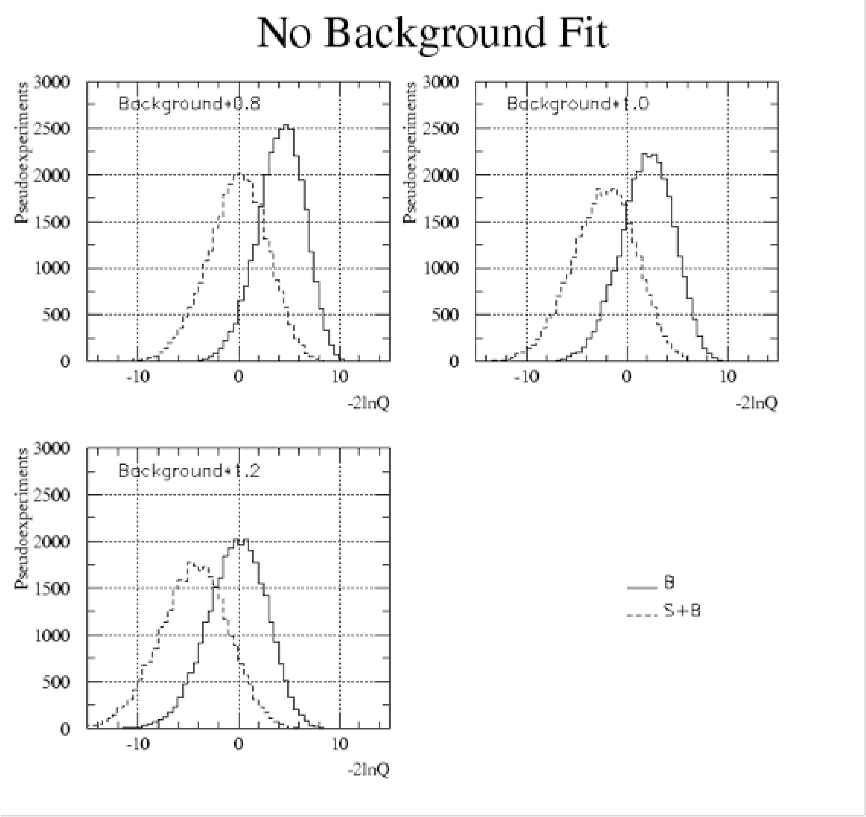
Fit to flat bg, Separate fit to flat bg + known signal shape.
The background rate is a nuisance parameter $\theta = \mathrm{b}$
Use fit signal and bg rates to calculate Q.
Fitting the signal is a separate option.







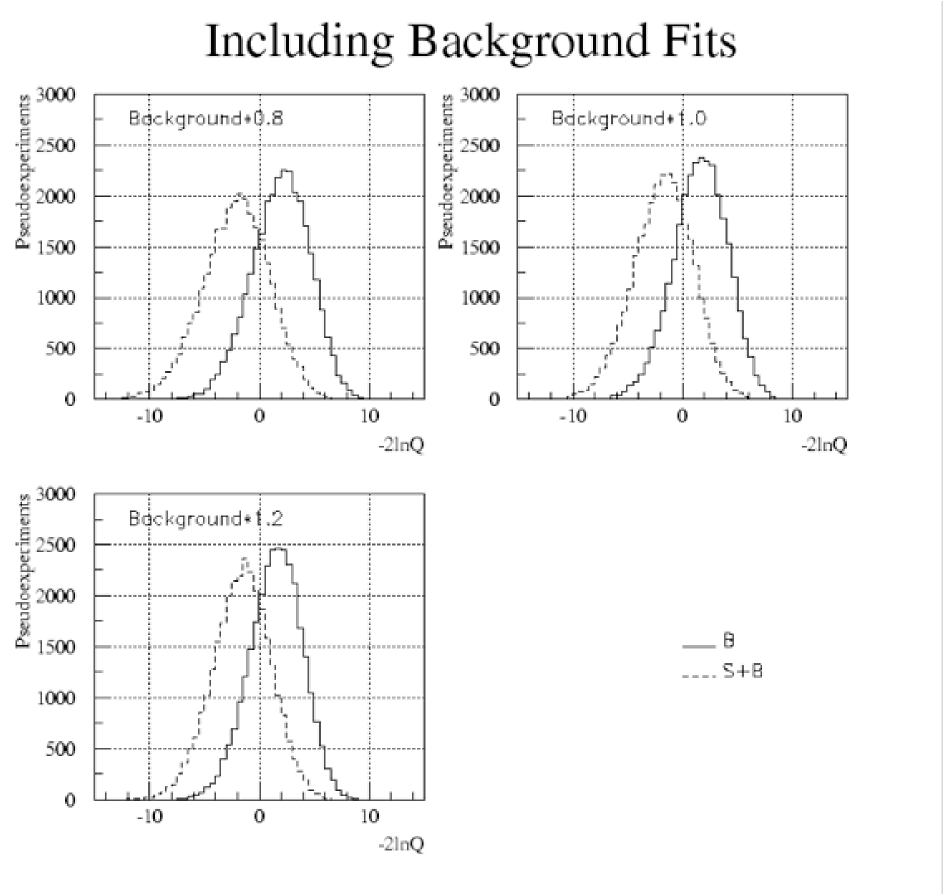
Fitting Nuisance Parameters to Reduce Sensitivity to Mismodeling



No Background Fit

Means of PDF’s of -2lnQ very sensitive to background rate estimation.

Still some sensitivity in PDF’s residual due to prob. of each outcome varies with bg estimate.



Including Background Fits

Some Comments on Fitting

- Fitting is an optimization step and is not needed for correctly handling systematic uncertainties on nuisance parameters.

  More on systematics later

- Some advocate just using -2lnQ with fits as the final step in quoting significance (Fisher, Rolke, Conrad, Lopez)

- Fits can "fail" -- MINUIT can give strange answers (often not MINUIT's fault). Good to explore distributions of possible fits, not just the one found in the data.

An Alternate Likelihood Ratio

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data} \mid H_1,\hat{\theta})}{L(\text{data} \mid H_0,\hat{\hat{\theta}})}\right)$$

Fit the signal freely in H_1. H_0 is then just a special case of H_1 (with s=0). Maximize over parameters of interest.

If we maximize the numerator, it will always then be at least as big as the denominator.

2lnQ will be distributed as a chisquared with one degree of freedom then -- Wilks's Theorem
(but -- need to check. MINUIT can give strange answers)

Expected p-values and Error Rates

- If H_0 is true, then the distribution of the p-value is uniform between 0 and 1
- If H_1 is true, then the distribution of p-values will be peaked towards smaller values (can be quite small if our sensitivity is large)
- We quote sensitivity as the median expected p-value if H_1 is true. Physicists say "sensitivity" -- statisticians use "power"
- Need to set a threshold for p-values to claim evidence or discovey (3σ and 5σ). These are the *error rates* e.g., 2.87E-7 is the error rate for false 5σ discoveries These are called "Type-I Errors" in stats jargon: rejecting H_0 when it's true.
- Can calculate probability of a 5σ discovery if H_1 is true -- spokespeople and lab directors like this.

Incorporating Systematic Uncertainties into the p-Value

Two plausible options:

"Supremum p-value"

Choose ranges of nuisance parameters for which the p-value is to be valid

Scan over space of nuisance parameters and calculate the p-value for each of those.

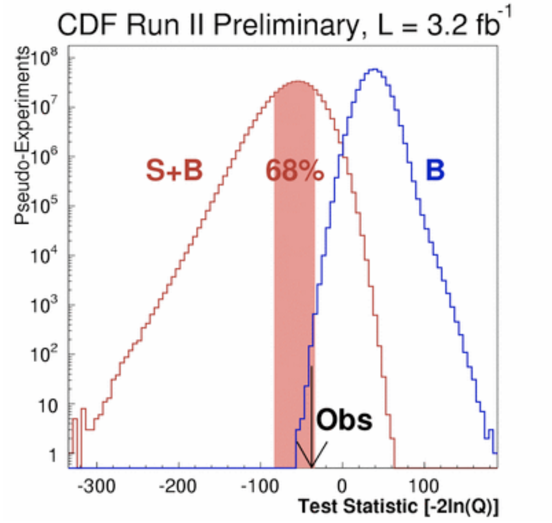
Take the largest (i.e., least significant, most "conservative") p-value.
"Frequentist" -- at least it's not Bayesian

"Prior Predictive p-value"

When evaluating the distribution of the test statistic, vary the nuisance parameters within their prior distributions. "Cousins and Highland"

Resulting p-values are no longer fully frequentist but are a mixture of Bayesian and Frequentist reasoning. In fact, adding statistical errors and systematic errors in quadrature is a mixture of Bayesian and Frequentist reasoning. But very popular.

Fitting and Fluctuating



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\mid s+b,\hat{\theta})}{L(\text{data}\mid b,\hat{\hat{\theta}})}\right)$$

- Monte Carlo pseudoexperiments are used to get p-values.
- Test statistic -2lnQ is not uncertain for the data.
- Distribution from which -2lnQ is drawn is uncertain!

- Nuisance parameter fits in numerator and denominator of -2lnQ **do not incorporate systematics into the result**.
  Example -- 1-bin search; all test statistics are equivalent to the event count, fit or no fit.

- Instead, we fluctuate the probabilities of getting each outcome since those are what we do not know. Each pseudoexperiment gets random values of nuisance parameters.

- Why fit at all? It’s an optimization. Fitting reduces sensitivity to the uncertain true values and the fluctuated values. For stability and speed, you can choose to fit a subset of nuisance parameters (the ones that are constrained by the data). Or do constrained or unconstrained fits, it’s your choice.

- If not using pseudoexperiments but using Wilk’s theorem, then the fits are important for correctness, not just optimality.

The Trials Factor

- Also called the "Look Elsewhere Effect"
- Bump-hunters are familiar with it.

What is the probability of an upward fluctuation as big as the one I saw *anywhere* in my histogram?

-- Lots of bins → Lots of chances at a false discovery

-- Approximation: Multiply smallest p-value by the number of "independent" models sought (not histogram bins!).

Bump hunters: roughly (histogram width)/(mass resolution)

Criticisms:

Adjusted p-value can now exceed unity!

What if histogram bins are empty?

What if we seek things that have been ruled out already?

The Trials Factor

More seriously, what to do if the p-value comes from a big combination of many channels each optimized at each m_H sought?

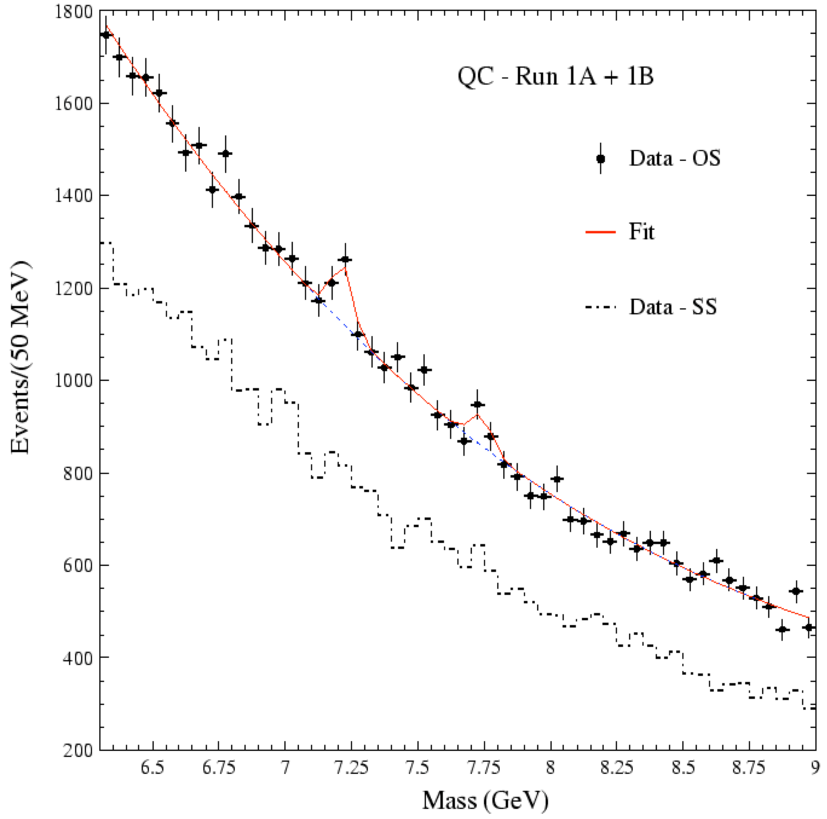
- Channels have different resolutions (or is resolution even the right word for a multivariate discriminant?
- Channels vary their weight in the combination as cross sections and branching ratios change with m_H

Proper treatment -- want a p-value of p-values!
(use the p-value as a test statistic)
Run pseudoexperiments and analyze each one at each m_H studied. Look for the distribution of smallest p-values.

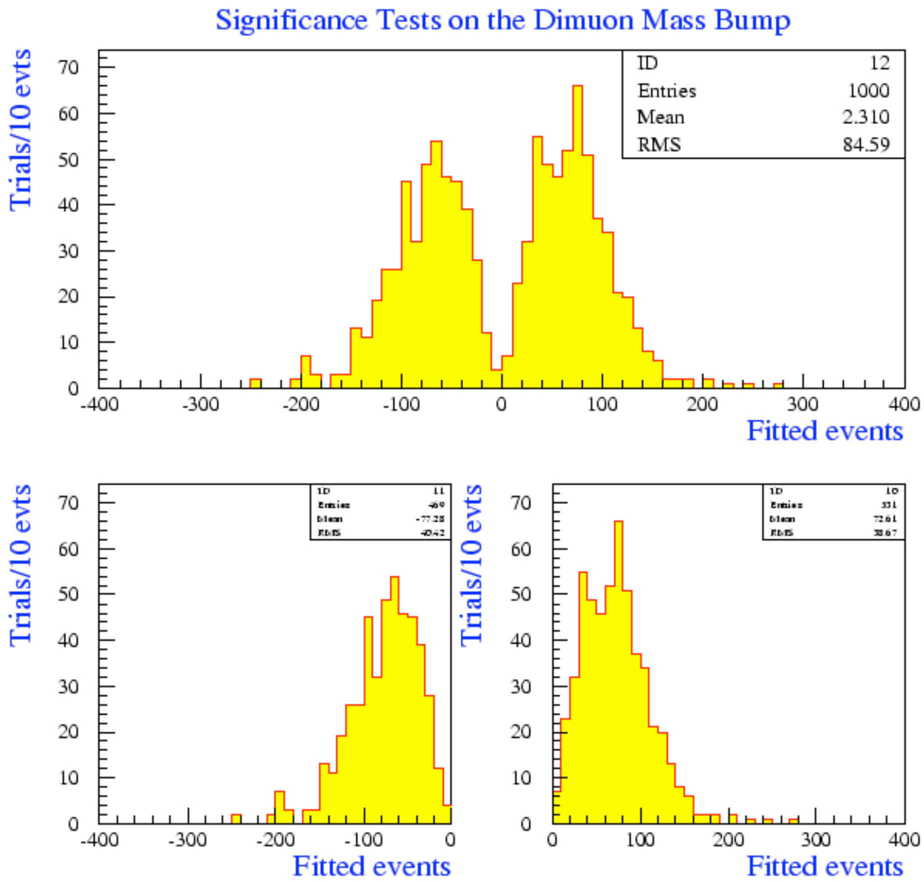
Next to impossible unless somehow analyzers supply how each pseudo-dataset looks at each test mass.

An internal CDF study that didn’t make it to prime time – dimuon mass spectrum with signal fit

(not enough PE’s)



249.7±60.9 events fit in bigger signal peak (4σ? No!)

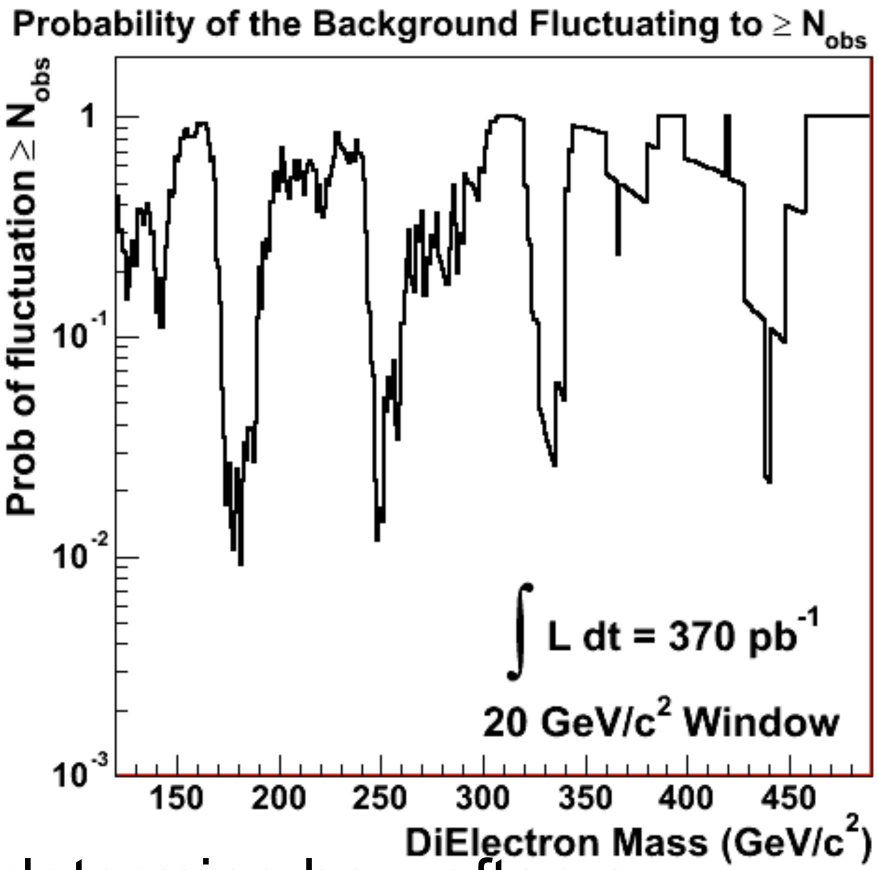


Null hypothesis pseudoexperiments with largest peak fit values

Looking Everywhere in a m_{ee} plot

- method:
  - scan along the mass spectrum in 1 GeV steps
  - at each point, work out prob for the bkg to fluctuate ≥ data in a window centred on that point
    - window size is 2 times the width of a Z' peak at that mass
  - sys. included by smearing with Gaussian with mean and sigma = bkg + bkg error
  - use pseudo experiements to determine how often a given probability will occur e.g. a prob ≤0.001 will occur somewhere 5-10% of the time

Single Pseudoexperiment



Aside -- Blind Analysis

- Fear of intentional or even unintentional biasing of results by experimenters modifying the analysis procedure after the data have been collected.

- Problem is bigger when event counts are small -- cuts can be designed around individual observed events.

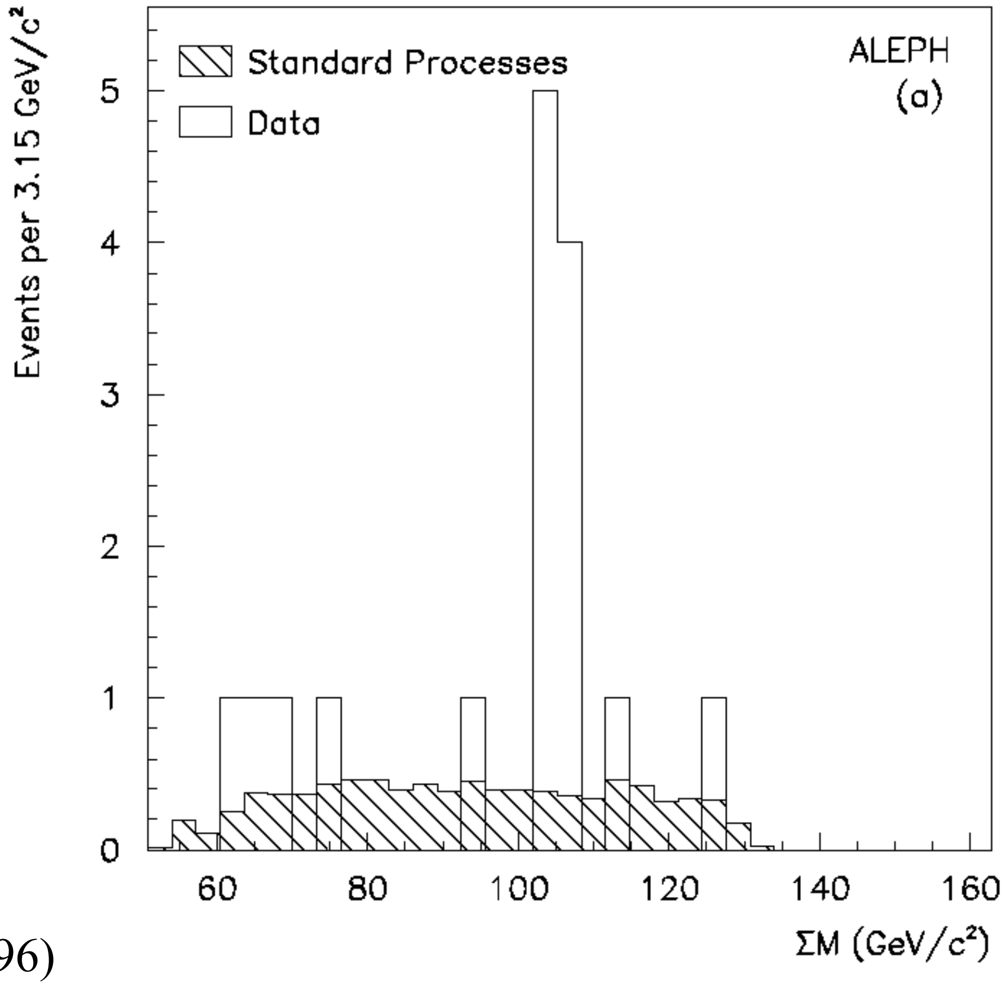
- Ideal case -- construct and optimize experiment before the experiment is run. Almost ideal -- just don't look at the data

- Hadron collider environment requires data calibration of backgrounds and efficiencies

- Often necessary to look at "control regions" ("sidebands") to do calibrations. Be careful not to look "inside the box" until analysis is finalized. Systematic errors too!

At Least they Explained what They Did

“the width of the bins is designed to correspond to twice the expected resolution ... and their origin is deliberately chosen to maximize the number of events found in any two consecutive bins”



ALEPH Collaboration, Z. Phys. C71, 179 (1996)

Dijet mass sum in $e^+e^- \rightarrow jjjj$

No Discovery and No Measurement? No Problem!

- Often we are just not sensitive enough (yet) to discover a particular new particle we're looking for, even if it's truly there.

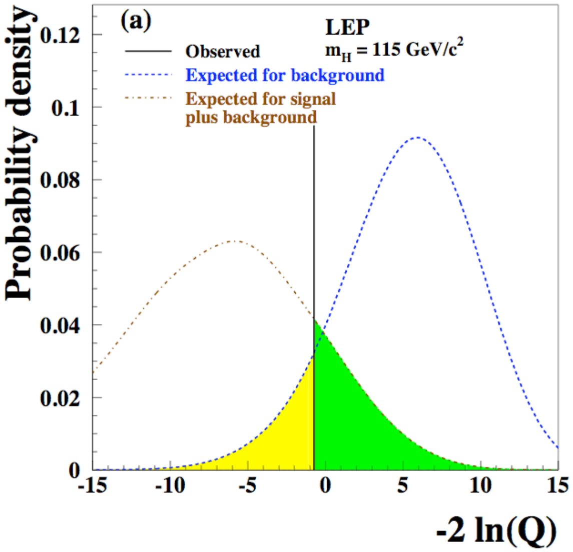
- Or we'd like to test a lot of models (each SUSY parameter choice is a model) and they can't all be true.

- It is our job as scientists to explain what we could have found had it been there. "How hard did you look?"

Strategy -- exclude models: set limits!

- Frequentist
- Semi-Frequentist
- Bayesian

CL_s Limits -- extension of the p-value argument



(apologies for the notation)

p-values:

$CL_b = P(-2\ln Q \geq -2\ln Q_{obs} |\ b\ \text{only})$

Green area = $CL_{s+b} = P(-2\ln Q \geq -2\ln Q_{obs}\ |\ s+b)$

Yellow area = "$1-CL_b$" = $P(-2\ln Q \leq -2\ln Q_{obs} | b\ \text{only})$

$CL_s \equiv CL_{s+b}/CL_b \geq CL_{s+b}$

Exclude at 95% CL if $CL_s<0.05$

Vary r until $CL_s=0.05$ to get r_{lim}

- Advantages:
  - Exclusion and Discovery p-values are consistent. Example -- a 2σ upward fluctuation of the data with respect to the background prediciton appears both in the limit and the p-value as such
  - Does not exclude where there is no sensitivity (big enough search region with small enough resolution and you get a 5% dusting of random exclusions with CL_{s+b})

A Simple Case -- CL_s in a Counting Search

-2lnQ is just a monotonic function of the observed number of events. In this case, more events is more "signal-like" (s+b>b). Not always the case

$CL_{s+b} = p(n \leq n_{obs}|s+b)$

Probability of s+b fluctuating downwards to n_{obs} or less (question: why not ask for equality?). "What is the chance of missing the signal this badly?

$CL_b = p(n \leq n_{obs}|b)$

Not quite 1-discovery p-value (equality flipped)

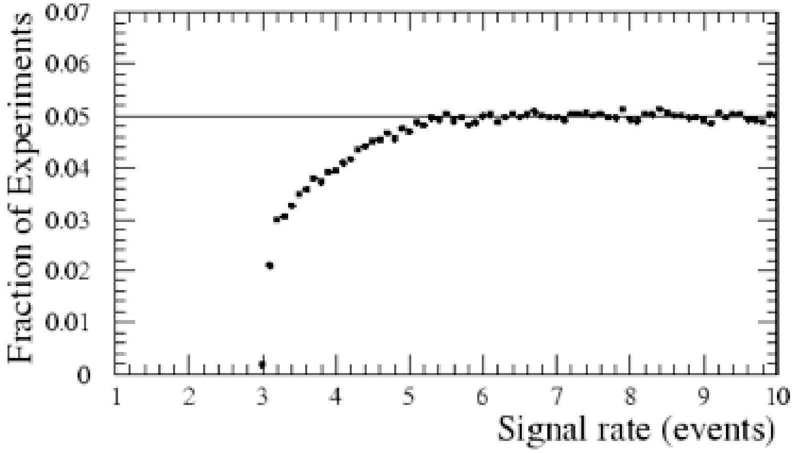
$CL_s = CL_{s+b}/CL_b$

Overcoverage on Exclusion

Coverage: The “false exclusion rate” should be no more than 1-Confidence Level

In this case, if a signal were truly there, we’d exclude it no more than 5% of the time. “Type-II Error rate” Excluding H_1 when it is true

Exact coverage: 5% error rate (at 95% CL)
Overcoverage: <5% error rate
Undercoverge: >5% error rate



T. Junk, NIM A434 (1999) 435.

Overcoverage introduced by the ratio $CL_s = CL_{s+b}/CL_b$
It’s the price we pay for not excluding what we have no sensitivity to.

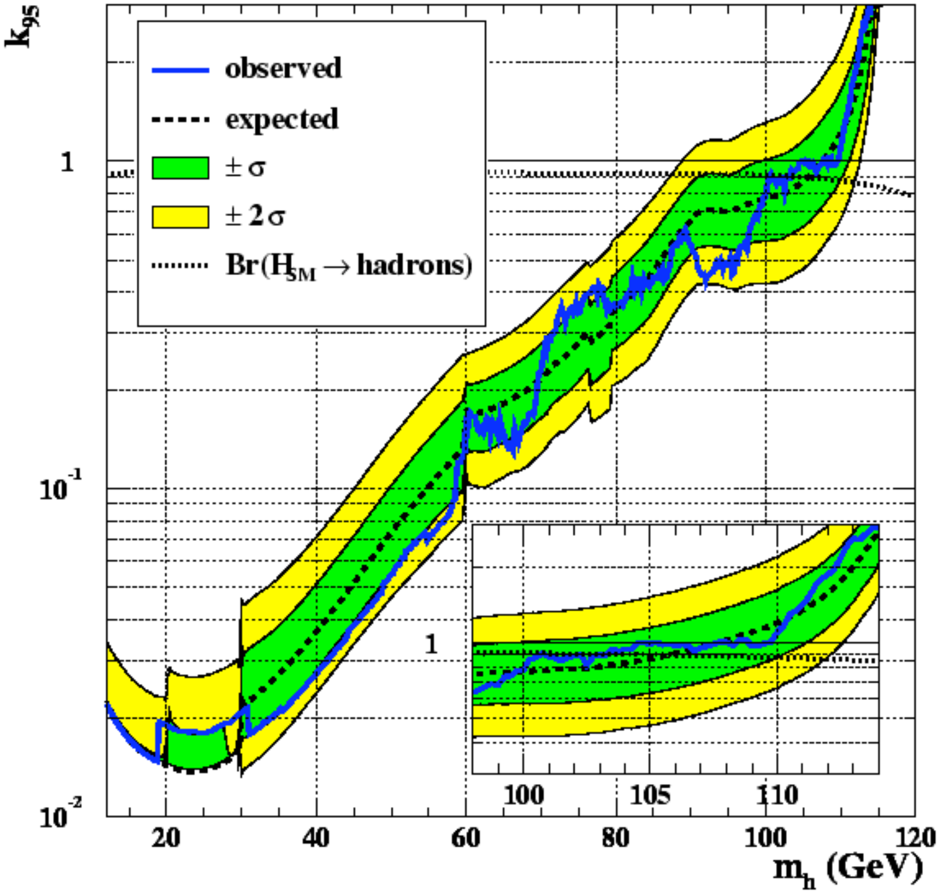
No similar penalty for the discovery p-value $1-CL_b$.

Different kinds of analyses switching on and off

OPAL’s flavor-independent hadronically-decaying Higgs boson search.

Two overlapping analyses: Can pick the one with the smallest median CL$_s$, or separate them into mutually exclusive sets.

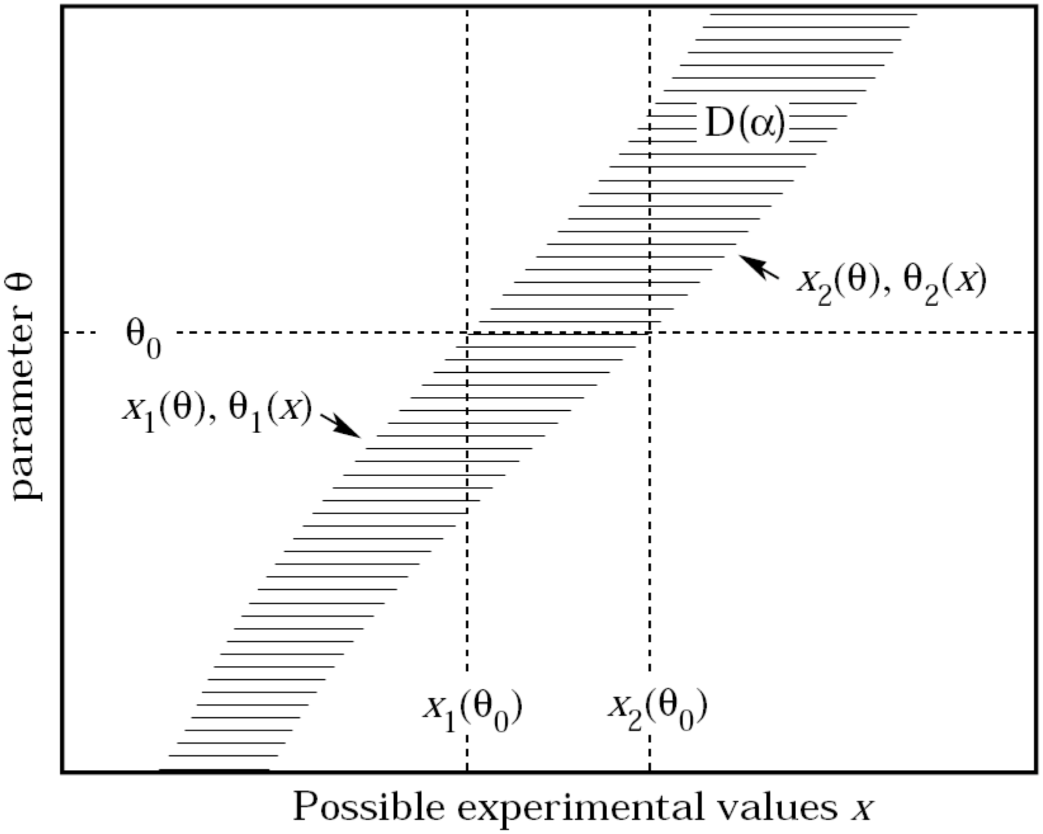
Important for SUSY Higgs searches.



The "Neyman Construction" of Frequentist Confidence Intervals

Essentially a
"calibration curve"

- Pick an observable *x* somehow related to the parameter *θ* you'd like to measure
- Figure out what distribution of observed x would be for each value of *θ* possible.
- Draw bands containing 68% (or 95% or whatever) of the outcomes
- Invert the relationship using the prescription on this page.


D(α)
$x_2(\theta)$, $\theta_2(x)$
θ_0
$x_1(\theta)$, $\theta_1(x)$
$x_1(\theta_0)$
$x_2(\theta_0)$
parameter θ
Possible experimental values *x*

A pathology: can get an
empty interval. But the error
rate has to be the specified one.
Imagine publishing that all branching ratios
between 0 and 1 are excluded at 95% CL.

Proper Coverage is Guaranteed!

A Special Case of Frequentist Confidence Intervals: Feldman-Cousins

Each horizontal band contains 68% of the expected outcomes (for 68% CL intervals)

But Neyman doesn't prescribe which 68% of the outcomes you need to take!

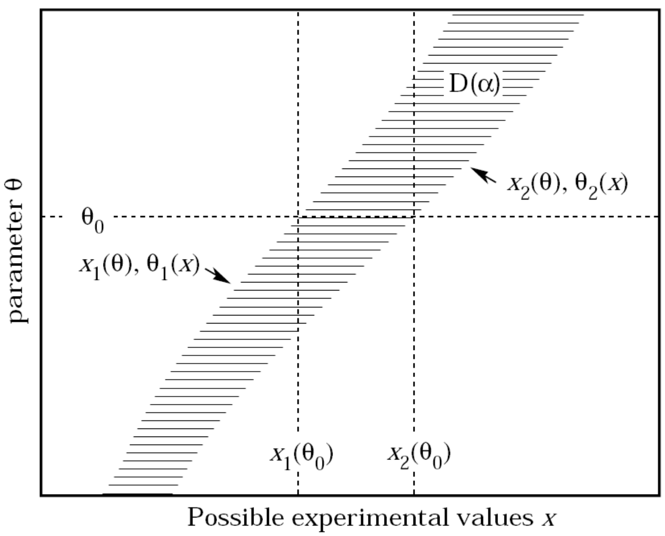
Take lowest x values: get lower limits.
Take highest x values: get upper limits.

Cousins and Feldman: Sort outcomes by the likelihood ratio.

$R = L(x|\theta)/L(x|\theta_{best})$

R=1 for all x for some θ.

Picks 1-sided or 2-sided intervals -- no flip-flopping between limits and 2-sided intervals.



D(α)
$x_2(\theta)$, $\theta_2(x)$
θ_0
$x_1(\theta)$, $\theta_1(x)$
$x_1(\theta_0)$
$x_2(\theta_0)$
parameter θ
Possible experimental values x

G. Feldman and R. Cousins, "A Unified approach to the classical statistical analysis of small signals" Phys.Rev.D57:3873-3889,1998. arXiv:physics/9711021

No empty intervals!

Some Properties of Frequentist Confidence Intervals

- Really just one: *coverage*. If the experiment is repeated many times, the intervals obtained will include the true value at the specified rate (say, 68% or 95%).

  Conversely, the rest of them (1-α) of them, must not contain the true value.

- But the interval obtained on a particular experiment may obviously be in the unlucky fraction. Intervals may lack credibility but still cover.

  Example: 68% of the intervals are from -∞ to +∞, and 32% of them are empty. Coverage is good, but power is terrible.

  FC solves some of these problems, but not all.
  Can get a 68% CL interval that spans the entire domain of θ.
  Imagine publishing that a branching ratio is between 0 and 1 at 68% CL.

  Still possible to exclude models to which there is no sensitivity.

  FC assumes model parameter space is complete -- one of the models in there is the truth. If you find it, you can rule out others even if we cannot test them directly.